Enumeration of RNA structures by Matrix Models

Graziano Vernizzi,[1] Henri Orland,[1] and A. Zee[2,3]

[1]*Service de Physique Théorique, CEA Saclay, 91191 Gif-sur-Yvette Cedex, France*
[2]*Department of Physics, University of California, Santa Barbara, CA 93106, USA*
[3]*Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106, USA*

We enumerate the number of RNA contact structures according to their genus, i.e. the topological character of their pseudoknots. By using a recently proposed matrix model formulation for the RNA folding problem, we obtain exact results for the simple case of an RNA molecule with an infinitely flexible backbone, in which any arbitrary pair of bases is allowed. We analyze the distribution of the genus of pseudoknots as a function of the total number of nucleotides along the phosphate-sugar backbone.

The prediction of foldings of single-stranded nucleic acids (like RNA molecules) is still a major open problem of molecular biology [1]. Several methods are available for the prediction and description of the folding process in various conditions. Most of them are statistical models (both at equilibrium and out-of equilibrium) that have roots in combinatorial problems. Although these models are much simpler than the energy based ones (and thus cannot provide thermodynamical predictions), they often provide exact analytical solutions that give important insights on the phase-space structure and the entropy. For those reasons the combinatorics of contact structures of biopolymers has received great attention over the past thirty years [2]. In the case of RNA-folding, a lot of attention has been paid to the combinatorics of contact structures that are planar (see e.g. [3] or [4] and references therein), but very little is known about non-planar structures (i.e. structures with pseudoknots). In this Letter we explore a very schematic model for RNA folding which allows for the exact enumeration of all contact structures with fixed genus. This model, which is based on a simpler one that was proposed earlier in [5, 6, 7, 8], may be relevant for studying the behaviour of non-planar contributions. The partition function is that of a chain of L nucleotides in three dimensions:

$$\mathcal{Z} = \int \prod_{k=1}^{L} d^3\mathbf{r}_k f(\{\mathbf{r}\}) Z_L(\{\mathbf{r}\}) \,, \qquad (1)$$

where $\mathbf{r}_k$ is the position vector in three dimensions of the k-th base, and $f(\{\mathbf{r}\})$ is a function which takes into account the geometry, the stiffness and the sterical constraints of the chain. The folding of the chain is caused by the hydrogen bonds that the bases can form. Since the hydrogen bonds saturate, a base can interact with only one other base at a time. The contribution from such interactions to the partition function is described by $Z_L(\{\mathbf{r}\})$:

$$Z_L(\{\mathbf{r}\}) = 1 + \sum_{i<j} V_{ij}(\mathbf{r}_{ij}) + \sum_{i<j<k<l} V_{ij}(\mathbf{r}_{ij}) V_{kl}(\mathbf{r}_{kl}) + \dots \,,$$

where $V_{ij}(\mathbf{r}_{ij}) = \exp(-\beta\varepsilon_{ij} v_{ij}(\mathbf{r}_{ij}))$ is the Boltzmann factor associated with the energy ε_{ij} of making a bond between the ith and the jth base at distance $\mathbf{r}_{ij}$. In this expression, $\beta = 1/T$ denotes the inverse temperature, and $v_{ij}(\mathbf{r}_{ij})$ represents the (short range) space dependent part of the interaction. To further simplify the model, we will assume that the chain is infinitely flexible and we will neglect all sterical constraints, so that any pairing of bases is assumed to be feasible. Therefore, we can neglect all spatial degrees of freedom and write

$$\mathcal{Z} = Z_L = 1 + \sum_{i<j} V_{ij} + \sum_{i<j<k<l} V_{ij} V_{kl} + \dots \,, \qquad (2)$$

where now $V_{ij} = \exp(-\beta\varepsilon_{ij})$. As shown in [5], each term in Z_L can be represented graphically by a suitable arc diagram. In such a representation the nucleotides are dots on an oriented horizontal line (which represents the RNA sugar backbone from the $5'$ end to the $3'$ end), and each base pair is drawn as an arc - above that line - between the two interacting bases. In real RNA, not all pairs of nucleotides can interact. For instance, two bases which are too close to each other along the backbone (say within a distance of 4 bases) cannot form a hydrogen bond since the backbone is not flexible enough. Moreover, for an RNA molecule one also usually assumes that only standard Watson-Crick pairs (A-U,C-G) and wobble pairs (G-U) are possible. These constraints greatly increase the difficulty of enumerating all possible structures that are allowed. Among the set of all possible structures, one defines *secondary structures* of an RNA molecule as all structures which are represented by planar arc diagrams (no crossing of arcs). When the diagrams are non planar, one says that the RNA molecule contains one or more *pseudoknots*. Structures with pseudoknots can be classified according to the topological character of the corresponding arc diagram [8]. Such a classification can be made more explicit directly in eq. (2), as explained in [5]. The main idea of [5] is to consider the following integral over matrices:

$$Z_L(N) = \frac{1}{A_L(N)} \int \prod_{k=1}^{L} d\varphi_k \, e^{-\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\,\mathrm{tr}(\varphi_i\varphi_j)} \times \frac{1}{N}\mathrm{tr}\prod_{l=1}^{L}(1+\varphi_l) \,. \qquad (3)$$

Here φ_i, $i = 1, \ldots, L$, are L independent $N \times N$ Hermitian matrices ($\varphi_i^+ = \varphi_i$) and $\prod_{l=1}^{L}(1+\varphi_l)$ is the ordered matrix product $(1+\varphi_1)(1+\varphi_2)\cdots(1+\varphi_L)$. The normalization factor is:

$$A_L(N) = \int \prod_{k=1}^{L} d\varphi_k e^{-\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\mathrm{tr}(\varphi_i\varphi_j)}\,, \qquad (4)$$

and V is the $L \times L$ symmetric matrix with elements V_{ij}. The integral in eq. (3) can be evaluated by using the Wick theorem. The result is a function of N which can be written as an asymptotic series at large N:

$$\begin{aligned} Z_L(N) \;&=\; 1 + \sum_{i<j} V_{ij} + \sum_{i<j<k<l} V_{ij}V_{kl} \\ &+\; \frac{1}{N^2}\sum_{i<j<k<l} V_{ik}V_{jl} + \ldots\,. \end{aligned} \qquad (5)$$

The relation with the expansion in eq. (2) is obvious. The two series coincide for $N = 1$, whereas for $N > 1$ the series in eq. (5) contains topological information. All the planar structures are given by the $O(1)$ term of eq. (5) and higher-order terms in $1/N^2$ correspond to RNA secondary structures with pseudoknots. The classification of pseudoknots induced by this expansion is reviewed in [8].

The most challenging problem in RNA-folding prediction is to find the structure with the lowest free energy. If one restricts the search to the set of secondary structures without pseudoknots, several fast algorithms are available [9]. However, when one includes the possibility of having pseudoknots, the problem is still open. An even simpler fundamental problem, namely the exact combinatorics of RNA structures with any pseudoknots, is unsolved. Results about the combinatorics of RNA secondary structures without pseudoknots or with very special classes of pseudoknots are available (e.g. [3, 4]), but the general case is still lacking. In this Letter we address precisely the problem of enumerating all secondary structures with pseudoknots.

In order to get exact results, we make a few additional simplifications. We assume that any possible pairing between nucleotides is allowed (independently of the type of nucleotides and from their distance along the chain) and that all the pairings may occur with the same probability. In other words, we assume that the matrix V_{ij} has all entries equal $v > 0$, i.e.:

$$V = \begin{pmatrix} v+a & v & \cdots & v \\ v & v+a & \cdots & v \\ \cdots & \cdots & \cdots & \cdots \\ v & v & \cdots & v+a \end{pmatrix}. \qquad (6)$$

The real number a has been added in order for V to be definite positive. Of course this addition is purely formal since $Z_L(N)$ does not depend on a, as one can easily see from eq. (5). In fact no diagonal term V_{ii} appears, as there are no self interaction diagrams. Even though the combinatorial problem in eq. (2) is now greatly simplified, it still keeps a lot of its topological interest. In fact, by means of the matrix integral in eq. (3) we can study the distribution of RNA structures with pseudoknots as a function of their topological character. Let us illustrate this point by a simple example for $L = 4$. In this case all possible contact structures are listed in Figure 1.

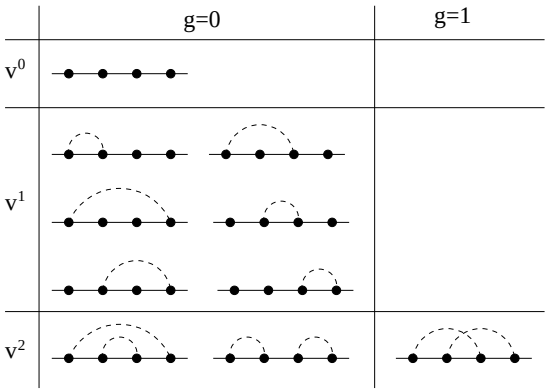

FIG. 1: All possible arc diagrams with $L = 4$. Diagrams with i arcs are associated to the power v^i, and g is the genus.

There is a total of ten possible arc diagrams, nine of which are planar and one which is not planar. The nine planar diagrams contain one diagram without arcs, six with one arc and two with two arcs. The same result can be directly obtained by computing the matrix integral in eq. (3). In fact, as we will show later in this Letter, the integral evaluates precisely to $Z_4(N) = 1+6v+2v^2+v^2/N^2$. Thus the coefficients of the asymptotic series have a direct topological interpretation, and that is the reason why the asymptotic $1/N^2$ expansion is usually referred as *topological expansion* [10]. Each term of the series gives the number of diagrams with fixed topological character: the first term represents planar diagrams, the second represents diagrams which can be drawn planarly on a surface with one handle (the torus), the third are diagrams that can be drawn planarly on a plane with two handles and so on. If we evaluate the integral in eq. (3) for any finite L and finite N, we will have an analytical control over the topology and the combinatorics of eq. (2). In the rest of the Letter, we will show explicitly how to compute the integral in eq. (3).

First, we note that by using a series of Hubbard-Stratanovich transformations, eq. (3) can be exactly simplified to:

$$Z_L(N) = \frac{1}{\tilde{A}(N)} \int d\sigma\, e^{-\frac{N}{2v}\,\mathrm{tr}\sigma^2}\, \frac{1}{N}\mathrm{tr}(1+\sigma)^L\,. \qquad (7)$$

We see that the original integration over the L matrices φ_k in eq. (3) has been reduced to an integration over a single $N \times N$ matrix σ. The similarity of eq. (7) and (3) is

obvious, and will be demonstrated in a future publication ([11]). Note that the regulator a drops out as long is it not zero. The normalization factor $\tilde{A}(N)$ is:

$$\tilde{A}(N) = \int d\sigma \, e^{-\frac{N}{2v}\,\mathrm{tr}\sigma^2} = \left(\frac{\pi v}{N}\right)^{\frac{N^2}{2}} 2^{\frac{N}{2}} \,. \qquad (8)$$

The Gaussian matrix integral in eq. (7) is straightforward. We introduce the spectral density of the matrix σ at finite N:

$$\rho_N(\lambda) \equiv \frac{1}{\tilde{A}(N)} \int d\sigma \, e^{-\frac{N}{2v}\,\mathrm{tr}\sigma^2} \frac{1}{N} \mathrm{tr}\, \delta(\lambda - \sigma) \,. \qquad (9)$$

By inserting the identity $1 = \int_{-\infty}^{+\infty} d\lambda \, \rho_N(\lambda)$ into eq. (7), we obtain:

$$Z_L(N) = \int_{-\infty}^{+\infty} d\lambda \, \rho_N(\lambda)(1+\lambda)^L \,. \qquad (10)$$

Thus the multi-dimensional integral of eq. (3) has been reduced to a one-dimensional integral. At this point it is convenient to study the exponential generating function of $Z_L(N)$:

$$G(t,N) \equiv \sum_{L=0}^{\infty} Z_L(N) \frac{t^L}{L!} = \int_{-\infty}^{+\infty} d\lambda \, \rho_N(\lambda) e^{t(1+\lambda)} \,. \qquad (11)$$

The explicit form of $\rho_N(\lambda)$ is a well known and classic result of Random Matrix Theory (see e.g. [12] or [13], and we use it in the form given in [14]):

$$\rho_N(\lambda) = \frac{e^{-\frac{N}{2v}\lambda^2}}{\sqrt{2\pi v N}} \sum_{k=0}^{N-1} \left(\begin{array}{c} N \\ k+1 \end{array} \right) \frac{H_{2k}(\lambda\sqrt{\frac{N}{2v}})}{2^k k!} \,, \qquad (12)$$

where $H_k(x)$ are the Hermite polynomials:

$$H_k(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2} \,. \qquad (13)$$

By inserting eq. (12) into eq. (11), one obtains:

$$G(t,N) = \frac{1}{N} \sum_{k=0}^{N-1} \left(\begin{array}{c} N \\ k+1 \end{array} \right) \frac{(t^2 v)^k}{k! N^k} e^{\frac{v\,t^2}{2N} + t} \,, \qquad (14)$$

where we have used the formula:

$$\int_{-\infty}^{+\infty} dx \, e^{-x^2 + xy} H_n(x) = y^n e^{y^2} \sqrt{\pi} \,. \qquad (15)$$

The sum in eq. (14) can be expressed as a generalized Laguerre polynomial:

$$L_N^{(1)}(z) = \sum_{k=0}^{N} \left(\begin{array}{c} N+1 \\ N-k \end{array} \right) \frac{(-z)^k}{k!} \,. \qquad (16)$$

We finally obtain:

$$G(t,N) \;=\; e^{\frac{v\,t^2}{2N} + t} \frac{1}{N} L_{N-1}^{(1)}\left(-\frac{v\,t^2}{N}\right) \,. \qquad (17)$$

From this exact result we can extract informations on all the coefficients $Z_L(N)$. The series expansion in t of $G(t,N)$ gives the first few coefficients $Z_L(N)$:

L	$Z_L(N)$
1	1
2	$1 + v$
3	$1 + 3v$
4	$1 + 6v + 2v^2 + v^2/N^2$
5	$1 + 10v + 10v^2 + 5v^2/N^2$
6	$1 + 15v + 30v^2 + 5v^3 + (15v^2 + 10v^3)/N^2$
7	$1 + 21v + 70v^2 + 35v^3 + (35v^2 + 70v^3)/N^2$
8	$1 + 28v + 140v^2 + 140v^3 + 14v^4 + (70v^2 + 280v^3 + 70v^4)/N^2 + 21v^4/N^4$

The meaning of these values is straightforward: the power of v is the number of arcs in the diagram, and the power of $1/N^2$ is the genus of the diagram. For instance when $L = 7$ there are 21 planar diagrams with one arc, and 35 diagrams on the torus (i.e. genus one closed oriented surface) with two arcs. The total number of diagrams for each fixed genus can be obtained by putting $v = 1$ (for instance, the total number of diagrams on the torus for $L = 6$ is 25). Analogously, the total number of diagrams, irrespective of the genus, can be obtained by putting $N = 1$ (for instance, the number of diagrams for $L = 4$ with 2 arcs is 3).

The general $1/N^2$ topological expansion of $Z_L(N)$ with $v = 1$ is:

$$Z_L(N) = \sum_{L=0}^{\infty} a_{L,g} \frac{1}{N^{2g}} \,, \qquad (18)$$

where the coefficients $a_{L,g}$ give exactly the number of diagrams at fixed length L and fixed genus g. From formula (17) and eq. (18) we recursively obtain all the coefficients $a_{L,g}$. Moreover, by normalizing each $a_{L,g}$ by the total number of diagrams at fixed L, i.e. by $\mathcal{N} \equiv Z_L(1)$, we can obtain the distribution of the number of diagrams. In Figure 2 we plot the distributions of diagrams as a function of L and g. We note the interesting feature that for any given $L >> 1$ most of the diagrams are not planar, and they have a genus close to a characteristic value $\langle g \rangle_L$. Such a value increases with L: we find numerically that it scales like $\langle g \rangle_L \sim 0.23L$, at large L. Also, for each fixed L there is a maximum possible value for g, namely $g \leq L/4$. Conversely a structure can have a genus g only if it has a length at least $L \geq 4g$.

It is important to note that even though the number of planar diagrams, $a_{L,0}$, is exactly the number of secondary structures without pseudoknots, and the number of diagrams on a torus, i.e. $a_{L,1}$ counts structures with one pseudoknot only, $a_{L,g}$ with $g \geq 2$ counts structures that contains either a single topologically complex pseudoknot, or several simple pseudoknots with small genus. For that reason, the concept of *irreducible pseudoknots* has been introduced in [5], and it would be of interest to study their distribution. The present analysis will be extended to the case of irreducible pseudoknots in a future

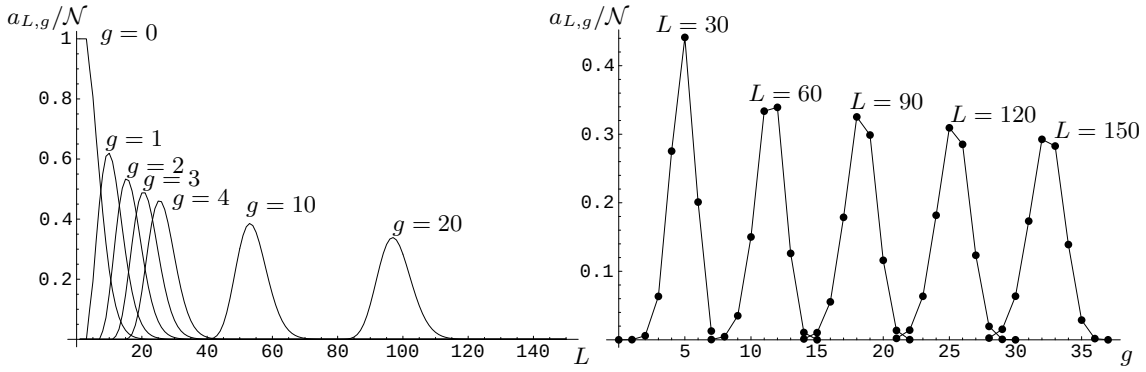

FIG. 2: On the left: the normalized number of diagrams $a_{L,g}/\mathcal{N}$ at fixed g as a function of L. On the right, the same quantity at fixed L as a function of g. In both cases we put $v = 1$.

publication, where we will also compute exact asymptotic behaviours for long sequences ([11]).

In this Letter, we have shown how one can compute the number of folded structures as a function of the length and of the genus of the RNA. This model is of course very schematic and oversimplified. It shows however that for a random RNA, the average topological character scales linearly with the length of the chain. As most wild RNA have an almost planar structure (with a genus $g \leq 2$), this implies that their sequences have been greatly designed by evolution in order to achieve this specificity.

Acknowledgments: We wish to thank G. Cicuta, W. Eaton, B. Eynard, P. Di Francesco, E. Guitter, S. Nonnenmacher, L. Molinari, for useful discussions. This work was supported in part by the National Science Foundation under grant number PHY 99-07949. GV acknowledges the support of the European Fellowship MEIF-CT-2003-501547.

[1] I. Tinoco Jr. and C. Bustamante, J. Mol. Biol. **293** (1999) 271.

[2] M.S. Waterman, Adv. Math. Suppl. Studies, **1** (1978) 167-212.

[3] I.L. Hofacker, P. Schuster and P.F. Stadler, Discr. Appl. Math. 88 (1998) 207-237.

[4] M.E. Nebel, Journal of Computational Biology 9/3 (2002) 541-574.

[5] H. Orland and A. Zee, Nucl. Phys. B [FS] **620** (2002) 456-476.

[6] M. Pillsbury, H. Orland and A. Zee, http://arXiv.org/physics/0207110.

[7] M. Pillsbury, J. A. Taylor, H. Orland and A. Zee, http://arXiv.org/cond-mat/0310505.

[8] G. Vernizzi, H. Orland and A. Zee, http://arxiv.org/abs/q-bio.BM/0405014.

[9] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker and D.H. Turner, Proc. Nat. Acad. Sci. 101 (2004) 7287-7292. See also I. L. Hofacker, Nucleic Acids Research, Vol. 31 No.13 (2003) 3429-3431.

[10] G. 't Hooft, Nucl. Phys. B **72** (1974) 461.

[11] G. Vernizzi, H. Orland and A. Zee, to be published.

[12] M.L. Mehta, Random Matrices, 2nd ed., Academic Press, New York, 1991.

[13] E. Brézin and A. Zee, Nucl. Phys. B **402** (1993) 613-627.

[14] G. Akemann, G.M. Cicuta, L. Molinari and G. Vernizzi, Phys.Rev.E 59 (1999) 1489-1497.